AD-A052 409 SPEECH COMMUNICATIONS RESEARCH LAB INC SANTA BARBARA CALIF F/G 9/2
ACOUSTIC/LINGUISTIC ASPECTS OF AUTOMATIC SPEECH RECOGNITION.(U)
MAR 78 D J BROAD, L L PFEIFER F44620-74-C-0034

UNCLASSIFIED AFOSR-TR-78-0692 NL

1 OF 1
ADA
052409

END
DATE
FILMED
5-78
DDC

18 19 **REPORT DOCUMENTATION PAGE**	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 78-0692 · 2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) 6 2 Acoustic/Linguistic Aspects of Automatic Speech Recognition.	5. TYPE OF REPORT & PERIOD COVERED Interim (1) Interim progress rept.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) 10 David J. Broad Larry L. Pfeifer	8. CONTRACT OR GRANT NUMBER(s) 15 F44620-74-C-0034
9. PERFORMING ORGANIZATION NAME AND ADDRESS Speech Communications Research Lab, Inc. 800 A Miramonte Drive Santa Barbara, CA 93109	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 16 17 61102F 2304/A2
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB, Washington, DC 20332	12. REPORT DATE 11 3 Mar [redacted] 78
	13. NUMBER OF PAGES 11 12 13 p.
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) 9 Interim progress rept. 1 Jan-31 Dec 77,	15. SECURITY CLASS. (of this report) UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved forpublic release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

This report describes:

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

~~During the 1977 contract year our work for the AFOSR has accomplished the following goals:~~ (1) Acquisition of a carefully labeled and segmented data base of connected speech for the testing of segmentation and phonetic identification algorithms. (2) Development and testing against the data base of a segmentation algorithm based on rate of spectral change and rms energy (3) Initiation of a study of inter-speaker vowel-format scaling based on a 2-dimensional constraint on a speaker's first three formant frequencies.

387 936

AFOSR-TR- 78-0692

INTERIM PROGRESS REPORT
OF RESEARCH ON
ACOUSTIC/LINGUISTIC ASPECTS OF AUTOMATIC SPEECH RECOGNITION

David J. Broad, Ph.D.
Larry L. Pfeifer, Ph.D.
Speech Communications Research Laboratory, Inc.
800 A Miramonte Drive
Santa Barbara, California

Contract F44620-74-C-0034

Air Force Office of Scientific Research

Reporting Period: January 1, 1977 - December 31, 1977

March 3, 1978

Contents

Page

1. Items of Progress

During the 1977 contract year our work for the Air Force Office of Scientific Research has accomplished the following goals:

1. Acquisition of a carefully labeled and segmented data base of connected speech for the testing of segmentation and phonetic identification algorithms.
2. Development and testing against the data base of a segmentation algorithm based on rate of spectral change and rms energy.
3. Initiation of a study of inter-speaker vowel-formant scaling based on a 2-dimensional constraint on a speaker's first three formant frequencies.

2. Description of Progress

2.1. Connected Speech Data Base

2.1.1. Motivation. One of the main bottlenecks for developing effective procedures for the processing and recognition of continuous speech, especially unconstrained conversational speech, is the acquisition of carefully labeled speech data to test algorithms for phonetic analysis. A part of the past year's effort has therefore been directed toward building up such a data base. This data base should be a valuable resource for the testing of any future algorithms for analyzing continuous speech.

To date, about 15 seconds of continuous speech has been analyzed according to the following procedures.

2.1.2. Marking Procedures. Using the Interactive Laboratory System (ILS) developed under previous support from AFOSR, a segment of speech of 1-3 minutes duration is stored on disk as a waveform bandlimited to 5 kHz and sampled at 10 kHz. The displayed waveform, formant frequencies, and rms energy are used together with repeated audio playback to assist the operator in making the best possible judgment for segmentation and transcription of each 100-frame interval (640 ms) of the waveform. Hard copies of the waveform and parameter display are preserved with their segmentation markers and phonetic transcriptions. At the same time, a label file is prepared which contains a greatly simplified form of the transcription. Each segment is marked as a V, S, N, C, or Z depending on its classification, respectively, as a vowel (V), sonorant constant (w, l, r, j) (S), nasal consonant (N), other consonant (C), or non-speech (silence or other non-speech) (Z). Ultimately, it would be desirable to encode the full phonetic transcription in the label file so that all the available phonetic information could be accessible to label-referenced algorithms. The labor involved for the present encoding system would make this not presently cost-effective.

2.1.3. Control on Subjectivity. Because even the best human transcriptions of connected speech contain some uncontrolled subjective factor, the above process is repeated on each speech segment by two transcribers working independently. After a transcription is completed by both workers, disagreements between the transcriptions are noted and, by working together, the transcribers then resolve most of the disagreements by discussion and re-examination of the data. Usually, agreement is easily achieved. There is, however, always some residual disagreement or uncertainty in difficult parts of the transcription. These are left as points of ambiguity in the final transcription.

2.1.4. Consistency. A comparison of the two transcriptions for the same 15 second interval showed that one experimenter transcribed and labelled 125 segments, while the other experimenter transcribed and labelled 135 segments. While the number of segments differed by 10, the number of discrepancies between the two transcriptions was 19.

It is found that the two transcribers are fairly consistent with each other in their placement of segment boundaries. A one- or two- frame discrepancy is not uncommon, and the average placement is consistent to within about 10 ms. Specifically, 33 percent of the boundaries are in perfect agreement, 67 percent of the boundaries are within 6.4 msec or less, and 83 percent of the boundaries are within 12.8 msec or less.

These figures, then, provide a useful guideline for evaluating automatic segmentation algorithms: their agreement with human transcription need not be better than the agreement of the humans with each other. Note that the figure of about 10 ms is of the same order as a single pitch period of a male voice; it might therefore be taken to represent a measure of inherent uncertainty of event timing in speech.

2.2. Segmentation Algorithm

2.2.1. Description. In order to build up a substantial data base for meaningful studies of conversational speech data, it is expected that automatic algorithms will be necessary for segmenting and labeling speech events. One such algorithm has been devised for the automatic detection of segment boundaries. This is accomplished by computing the spectral variance of the speech signal as a function of time. The variance function tends to have local maxima in the transition region between sounds and local minima in sounds which can have steady-state characteristics. Thus a potential segment boundary is placed at the location of peaks in the variance function. All potential boundary markers are displayed on the graphics terminal for visual verification, but the operator must still identify and label the marked segments. This boundary detection algorithm is speaker-independent and operates reliably on unconstrained speech.

2.2.2. Human vs. Machine Marking. The performance of the automatic algorithm was evaluated at one level by comparing the location of vowel boundaries placed by the machine versus those placed by one of the transcribers. It was found that there was complete agreement on 33 percent of the initial vowel boundaries. Furthermore, 66 percent of the initial vowel boundaries and 52 percent of the final vowel boundaries were within 6.4 msec of each other. Also, 81 percent of the initial vowel boundaries and 81 percent of the final vowel boundaries were within 12.8 msec of each other. Note that this is very close to the level of agreement between the two transcribers.

These results are very encouraging and demonstrate the effectiveness of the spectral variance function in locating vowel boundaries. It is expected that this algorithm could be the foundation for a totally automatic segmentation and labeling process.

2.3. Vowel Scaling Study

2.3.1. Background. As reported by Broad and Wakita (1977), a large sampling of a given speaker's first three vowel formant frequencies cluster near a 2-part 2-dimensional surface of the form:

$$\alpha_1 F_1 + \alpha_2 F_2 + \alpha_3 F_3 + \alpha_4 = 0 \quad \text{for } F_2 > \Theta_2 \qquad (1)$$

$$\beta_1 F_1 + \beta_2 F_2 + \beta_3 F_3 + \beta_4 = 0 \quad \text{otherwise.}$$

They noted that this constraint on the formant frequencies had important implications for the problem of inter-speaker formant scaling. In particular, the hypothesis of uniform scaling would imply that any speaker's distribution of vowel formant frequencies should have the form (1) modified only by replacing α_4 and β_4 with α_4/k and β_4/k, where k is a speker-dependent scaling factor related to the average vocal tract length.

2.3.2. New Results. To study this aspect of inter-speaker scaling, we have been collecting data on the vowels of 4 new speakers. The study is to include 2-6 additional speakers beyond these. The data collected so far on the 4 speakers as reduced to the form (1) are shown in Table I together with those of the speaker studied by Broad and Wakita. These

Table I. Parameters of the representation (1) for four speakers compared to those of the speaker (F2) studied by Broad and Wakita (1977). N is the number of samples used for each speaker, and σ is the rms distance of the samples from the representation (1). The F_2 threshold for discriminating front and back vowels is Θ_2 in (1). (F = female; M = male).

				Front Vowels				Back Vowels			
Speaker	N	σ (Hz)	Θ_2 (Hz)	α_1	α_2	α_3	α_4	β_1	β_2	β_3	β_4
F1	459	85	1,500	.59	.71	-.38	-816	.62	-.63	-.46	1,912
F2	778	86	1,710	.63	.60	-.49	-366	.69	-.53	-.50	1,569
M1	1,399	85	1,250	.89	.34	-.31	-234	.81	-.48	-.33	1,398
M2	1,763	72	1,500	.76	.58	-.28	-732	.38	-.43	-.92	2,586
M3	380	71	1,300	.94	.32	-.10	-787	.75	-.45	-.48	1,270

preliminary results suggest the following tentative conclusions:

(1) The two-plane form (1) provides a satisfactory description of vowel formant frequencies across speakers, inasmuch as the rms spread of the data about these planes is about the same for all the speakers.

(2) The various speakers' representations are similar at least to the extent that all their direction cosines ($\alpha_1 - \alpha_3$, $\beta_1 - \beta_3$) have the same respective signs, as do the average offsets in Hz (α_4 and β_4).

(3) The α's and β's are quite variable from speaker to speaker, suggesting that the hypothesis of uniform scaling cannot be supported. This will be determined conclusively from a more formal analysis of the complete data set.

Whether some modified form of uniform scaling can be made to work remains to be seen.

Reference

D.J. Broad and H. Wakita, Piecewise-Planar Representation of Vowel Formant Frequencies, *The Journal of the Acoustical Society of America*, Volume 62, Number 6, December, 1977, pp. 1467-1473.

3. Publications

A revision of the following paper has been re-submitted to the *IEEE Transactions on Acoustics, Speech, and Signal Processing*:

L. L. Pfeifer, An Interactive Laboratory System for Research in Speech and Signal Processing.

The following manuscript is near completion and will be submitted to the same journal:

L. L. Pfeifer, Methodologies for Acoustic Studies of Vowels in Conversational Speech.